Learning Domain Invariant Representations for Generalizable Person Re-Identification

Yi-Fan Zhang[1], Hanlin Zhang[2], Zhang Zhang[1], Da Li[1], Zhen Jia[1], Liang Wang[1], Tieniu Tan[1]
[1]Institute of Automation, Chinese Academy of Sciences [2]Carnegie Mellon university

Abstract

Generalizable person Re-Identification (ReID) has attracted growing attention in recent computer vision community, as it offers ready-to-use ReID models without the need for model retraining in new environments. In this work, we introduce causality into person ReID and propose a novel generalizable framework, named Domain Invariant Representations for generalizable person Re-Identification (DIR-ReID). We assume the data generation process is controlled by two sets of factors, i.e. *identity-specific factors containing identity related cues, and domain-specific factors describing other scene-related information which cause distribution shifts across domains. With the assumption above, a novel **M**ulti-**D**omain **D**isentangled **A**dversarial **N**etwork (MDDAN) is designed to disentangle these two sets of factors. Furthermore, a **C**ausal **D**ata **A**ugmentation (CDA) block is proposed to perform feature-level data augmentation for better domain-invariant representations, which can be explained as interventions on latent factors from a causal learning perspective. Extensive experiments have been conducted, showing that DIR-ReID outperforms state-of-the-art methods on large-scale domain generalization (DG) ReID benchmarks. Moreover, a theoretical analysis is provided for a better understanding of our method.*

1. Introduction

Person Re-IDentification (ReID) aims at matching person images of the same identity across multiple camera views. In earlier works, most ReID methods are trained and tested on the same dataset, termed fully-supervised methods [80, 83], or tested on new target domains different from the training one, termed domain adaptation (DA) methods [41, 52]. Although recent fully-supervised methods have achieved remarkable performance, they tend to fail catastrophically when tested in out-of-distribution (OOD) settings. Figure 1 illustrates the fragility of two representative fully-supervised works, *i.e.*, the DG-Net [83] and ISGAN [16], which get very high rank-1 accuracies above $94\%, 95\%$ respectively when tested on Market1501 dataset. However, the rank-1 accuracies drop to 18.7% and 27.8% respectively when directly tested on the dataset of GRID, which suggests the weak extrapolation capability and poor robustness of fully-supervised methods. We further train these two models over multiple source domains (the details of sources are in Section 4.1). However, the even worse performance are obtained, which indicate the challenge of cross-domain ReID. To tackle these problems, a number of DA methods have been proposed to mitigate the domain gap without the need for extensive manual annotations for new camera settings. However, as the deployment information of new cameras is unknown, large amounts of unlabeled data are required for DA retraining. These problems severely hinder real-world applications of current person ReID systems.

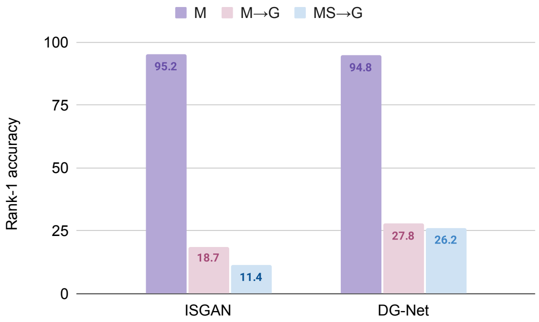


Figure 1. Performance of two representative models. M: Train and test on Market1501. M→G: Trained on Market1501 and tested on GRID . MS→G: Trained on multi-source datasets and tested on GRID.

To tackle the above challenges, we focus on a more realistic and practical setting: generalizable person ReID, where the model is trained on multiple large-scale public datasets forming source domains and tested on unseen domains. The generalizable person ReID is originally formulated as a problem of domain generalization (DG) [59], which is more practicable than the traditional ReID paradigm since the ready-made models can work on any new settings without repeating the whole process of data collection, annotation, and model updating.

Existing methods mainly follow a meta-learning pipeline [59] or utilize domain-specific heuristics [28, 62]. Different from them, we address the generalized person ReID problem based on a disentangled representation learning framework, where the domain information of various data sources is explicitly embedded and further used for augmenting unseen data samples to improve the disentanglement.

Inspired by the observation that human beings can easily distinguish the same person identities with their appearances, body shapes, or accessories from various visual contexts or imaging conditions such as backgrounds, viewing angles, illuminations, *etc*. Based on this intuition, we first present a causal view on the generalizable person ReID. With this view, we use the structural causal model (SCM) to provide insights for the poor generalization of traditional ReID models when applied to unseen domains. The assumption behind this is that person images are affected by two sets of latent random variables, *i.e*. identity-specific factors, and domain-specific factors. Thus, a desirable generalizable person ReID model should be able to identify these two kinds of factors in a disentangled latent space, leading to invariant perceptions of person identity across various data sources.

Based on the above analysis, firstly, a Multi-Domain Dual Adversarial Network (MDDAN) is proposed to jointly learn two encoders for embedding identity-specific and domain-specific factors from multiple public ReID datasets, where the adversarial learning principle is adopted to exclude the domain (identity) related information from the embedded identity (domain) specific representations. Then, motivated by the great benefit of data augmentation to DG [25, 77, 24], we further exploit the advantages of the disentangled representations with a Causal Data Augmentation (CDA) block. Here, a feature-level data augmentation is performed by mixing one identity-specific feature with various domain-specific features, without the need for image manipulation or manual data collection. The intuition behind the CDA block can be explained as interventions on the identity-specific and domain-specific factors from a causal learning perspective. In this way, an identity-preserving loss on the augmented data helps to obtain causal invariances across various data domains. These two components (MDDAN and CDA) are integrated as a whole system, termed DIR-ReID.

To sum up, the contributions of our work can be summarized as:

- For the first time, a causal perspective on the analysis of generalizable person ReID is introduced, based on which we claim that the essential problem of generalizable person ReID is how to disentangle identity-specific factors and domain-specific factors for improving robustness to spurious correlations;
- A disentangled representation learning approach named DIR-ReID is proposed, where a MDDAN block is adopted to identify identity-specific and domain-specific factors from multiple data sources. Then a CDA block is adopted for data augmentation as causal interventions. Mathematical analysis proves the characteristics of our method;
- Comprehensive experiments are conducted to demonstrate the effectiveness of the learned disentangled representations. Our method achieves superior performance in comparison with State-Of-The-Art (SOTA) methods on large-scale generalizable person ReID benchmarks.

2. Related Work

Single-domain Person ReID Existing works of single-domain person ReID (*i.e*. supervised person ReID) usually depend on the assumption that in training and testing data are independent and identically distributed (i.i.d.). They usually design or learn discriminative features [72, 30, 43] and develop efficient metrics [31, 35, 70]. With the rapid development of deep Convolutional Neural Networks (CNNs), single-domain person ReID has achieved great progress. Some of the CNN-based methods introduce human parts [32, 61], poses [79], and masks [58] to improve the robustness of extracted features. And some other methods use deep metric learning to learn appropriate similarity measurement [14, 11]. Despite the encouraging performance under the single-domain setup, current fully-supervised models for person ReID degrade significantly when deployed to an unseen domain.

Cross-domain Person ReID Unsupervised Domain Adaptation (UDA) has achieved great progress [50] and is commonly adopted for cross-domain person ReID to achieve good performance on unlabeled target domain with both the labeled source data and unlabeled target data. The UDA-based ReID methods usually attempt to transfer the knowledge from source domains to the target one depending on target-domain images [13, 41, 27], features [64] or metrics [49]. Another group of UDA-based methods [17, 18, 20, 75] propose to explore hard or soft pseudo labels in unlabeled target domain using its data distribution geometry. Though UDA-based methods improve the performance of cross-domain ReID to a certain extent, most of them require a large amount of unlabeled target data for adaptation.

Generalizable Person ReID Recently, Generalizable person ReID methods [59, 28, 29, 36, 67] are proposed to learn a model that can generalize to unseen domains without target data and the adaptation. Jia *et al*. [28] learn the domain-invariant features by integrating the Instance

Normalization (IN) into the network to filter out style factors. Jin *et al.* [29] extend the work [28] by restituting the identity-relevant information to network to ensure the model discrimination. Lin *et al.* [36] propose a feature generalization mechanism by integrating the adversarial auto-encoder and Maximum Mean Discrepancy (MMD) alignment. Song *et al.* [59] propose a Domain-Invariant Mapping Network (DIMN) following the meta-learning pipeline and maintaining a shared memory bank for training datasets. Wang *et al.* [67] improve the generalizability of learned model by synthesizing a large-scale person ReID dataset. Different from the above methods, our work addresses the DG person ReID via domain invariant representation learning..

Domain Generalization Domain/Out-of-distribution generalization [24] aims to learn representations $\Phi(X)$ that is invariant across environments $\mathcal{E}$ so that model can well extrapolate in unseen environments. The problem can be formulated as $\min_{\Phi} \max_{e \in \mathcal{E}} \mathbb{E}[l(y, \Phi(x)) \mid E = e]$. Representative approaches like IRM [2] and its variants [1] are proposed to tackle this challenge. However, IRM entails challenging bi-level optimization and would fail catastrophically unless the test data are sufficiently similar to the training distribution [54]. Data augmentation is another effective strategy to address DG [77] but it heavily depends on domain-specific expertise. To address those challenges, we proposed DIR-ReID which can be seen as an effective way for disentanglement and data augmentation.

Causality for CV Causal Representation Learning [57] combines machine learning and causal inference, which has attracted increasingly attention and become a potential direction towards human-level intelligence. Simultaneously, there is a growing number of computer vision tasks that benefit from causality [3, 42, 44, 53, 51, 39]. Most of them focus on causal effects: disentangling the desired model effects [7], and modularizing reusable features that generalize well [47]. Recently, causal intervention is also introduced into computer vision [66, 71, 63]. Specifically, CONTA [15] removes the confounding bias in image-level classification by backdoor adjustment and thus provides better pseudo-masks as ground-truth for the subsequent segmentation model. IFSL [74] believes that pre-training is a confounder hurting the few-shot learning performance. Specifically, they propose a SCM in the process of FSL and then develop three practical implementations based on the backdoor adjustment. We also adopt the SCM of Pearl [48] to model the causal effects of person ReID. However, our implementation is not based on frontdoor/backdoor adjustment and our target is not to infer the marginal distribution. We use causality to provide the explanations why traditional methods work poorly on unseen domains and further propose methods to tackle the problem.

3. Learning Disentangled and Invariant Representations

In this section, we first introduce the proposed SCM to illustrate the motivation to disentangle identity-specific factors and domain-specific factors for improving the robustness of generalizable person ReID to domain variations [44, 8, 83]. Then, a DIR-ReID model is proposed to learn the disentanglement of identity-specific factors and domain-specific factors from multiple public ReID datasets. Finally, a theoretical analysis is taken for a better understanding of our method.

3.1. SCM for Generalizable Person ReID

Inspired by current research of harnessing causality in machine learning [3], we propose a SCM to analyze the disentanglement and generalization in person ReID models.

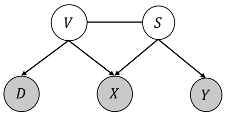


Figure 2. Graphical representation of person ReID methods. X: pedestrian images, S, V: identity-specific and domain-specific factors, Y, D: pedestrian and domain labels. Gray circles denote observable variables.

Following the causal models in [8, 60], we use the SCM (in Figure 2) to describe the causal relationships between person images, data domains and person identities, where X, Y and D denote the observable variables of pedestrian images, identity labels and domain labels, S and V are the latent variables indicating identity-specific and domain-specific factors respectively. As shown in the model, there are four kinds of causal relationships as follows:

$S, V \rightarrow X$. One person image is determined by its identity-specific factors, *e.g.*, body shapes, clothes, and accessories, as well as domain-specific factors, *e.g.*, backgrounds, illuminations, and viewpoints.

$S \rightarrow Y$. Identity-specific factors S directly cause Y, which means the person identities are only determined by their identity-specific information, such as clothing styles, body shapes, *etc.*, instead of the backgrounds, illuminations and camera viewpoints.

$V \rightarrow D$. Domain-specific factors are the direct causes of D, namely images captured by a multi-camera network (single domain) likely share similar illumination conditions, indoor/outdoor backgrounds.

$S - V$. The undirected $S - V$ edge indicates the spurious correlations between S and V in the real world, which expresses the indirect influences of V to Y (S to D). For example, pedestrians mostly appear in campus in the CUHK03 while in a busy underground station in the GRID.

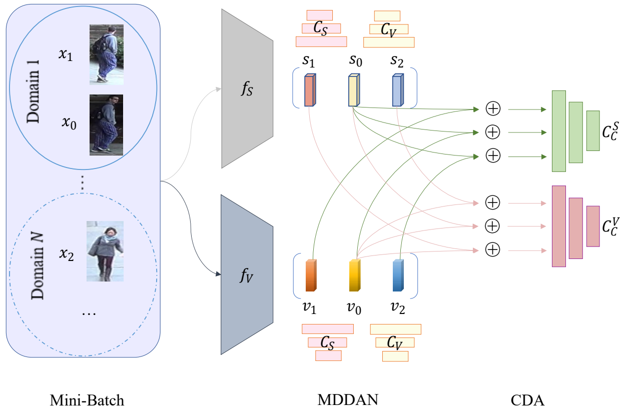


Figure 3. **Schematic description of the proposed approach.** x_0, x_1, x_2: Three images from one mini-batch. f_S, f_V: encoders of identity-specific and domain-specific factors. s, v: identity-specific and domain-specific factors represented in latent space. $\mathcal{C}_S, \mathcal{C}_V, \mathcal{C}_C$: classifiers for identity-specific factors, domain-specific factors, and concatenated vectors. $\oplus$: concatenation of two latent vectors.

So the pedestrian identities and their backgrounds are spuriously correlated.

Thanks to the causal structure, we can explain why V confounds the person ReID process: even if some pixels in X (which belong to domain-specific factors) have nothing to do with Y, the path $V - S \rightarrow Y$ can still correlate V and Y, which misleads the identity classifier to focus on these domain-specific factors. That is why the traditional supervised methods have poor generalization ability: when trained on a single domain, identity-specific factors and domain-specific factors are highly entangled in the feature space, while the spurious correlations learn from a single training domain do not hold in cross-domain tests. Thus, it is desirable to reduce such spurious correlations to improve the generalization capability of the learned representations. However, it is almost impossible to get an ideal dataset[1] covering all kinds of domain-specific factors. In this paper, we will explore and exploit the potentials of existing large-scale public datasets to learn identity-domain disentangled representations for generalizable person ReID.

3.2. The DIR-ReID Framework

Notations and Problem Formulation. For generalized person ReID, we have access to G labeled datasets $\mathcal{D} = \{\mathcal{D}_g\}_{g=1}^{G}$. Each dataset $\mathcal{D}_g = \{(x_i, y_i, d_i)\}_{i=1}^{N_g}$, where N_g is the number of images in $\mathcal{D}_g$ and (x_i, y_i, d_i) is the image, identity label and domain label for the image respectively. In the training phase, we train a DG model using $N = \sum_{g=1}^{G} N_g$ aggregated image-label pairs with $M = \sum_{g=1}^{G} M_g$ identities from all source domains. In the testing phase, we perform a retrieval task on unseen target domains without additional model update.

As analyzed in Section 3.1, the essential is how to get the disentanglement of S and V for a domain invariant representation. Here, we propose the DIR-ReID to tackle the challenging problem. DIR-ReID consists of three blocks, as shown in Figure 3.

(i) Identity adversarial learning block, which is a domain-agnostic, identity-aware encoder f_S to obtain identity-specific factors.

(ii) Domain adversarial learning block, which is a domain-aware, identity-agnostic encoder f_V to identify domain-specific factors.

(iii) Causal data augmentation block, which adopts data augmentation from a causal view in SCM for learning invariant representations.

Components (i) and (ii) are named MDDAN. As shown in figure 3, the overall process includes feeding images x_0, x_1, x_2 randomly selected from N source domain. into identity-specific and domain-specific encoders to get disentangled representations s_0, s_1, s_2 and v_0, v_1, v_2 via adversarial learning. Then causal data augmentation is performed based on these representations to train the encoders f_S and f_V for learning invariant representations.

Identity Adversarial Learning Block. An identity-aware encoder f_S is adopted to extract identity-specific factors. Then, a classifier $\mathcal{C}_s$ is used to identify the ID label for a given person image x_i. The cross-entropy loss with label smoothing [45] is calculated for training the encoder f_S, which is defined as:

$$\mathcal{L}_{id}^{s} = -\sum_{i=1}^{N} \log P\left(Y = y_i | \mathcal{C}_S\left(f_S\left(x_i\right)\right)\right) \tag{1}$$

where θ_S is the parameters of f_S, $\mathcal{C}_S$ is the identity classifier. y_i is the labeled person identity of x_i.

To exclude all the domain information from identity-specific factors, a domain classifier $\mathcal{C}_V$ is adopted for adversarial learning. One promising way is using gradient reversal layer (GRL) technique in DANN [19] to train the encoder f_S and classifier $\mathcal{C}_v$ simultaneously, where a misclassification loss is adopted to enforce an image not to be classified into its true domain class. However, "misclassification" does not mean "indistinguishability". Intuitively, "indistinguishability" means classifying one image into all the domains equiprobably. Hence, we adopt a loss of maximum entropy (minimization of negative entropy [8]) termed the domain-indistinguishability loss as follows:

$$\mathcal{L}_{indis}^{dom} = -\mathcal{H}(D|S)), \tag{2}$$

[1] An ideal dataset has images of every pedestrian in all backgrounds and illumination conditions, which is unbiased and has few spurious correlations. In such a case, traditional deep neural network (DNN)-based methods will not consider these domain-specific factors as identity-related, and thus the spurious correlations between S, V is removed.

where $\mathcal{H}$ is the entropy to measure the uncertainty of predicted domain class given identity-specific factors, *i.e.*,

$$\mathcal{H}(D|S)) = -\sum_{i=1}^{N}\sum_{g=1}^{G} P\left(D = g|\mathcal{C}_V\left(f_S\left(x_i\right)\right)\right) \cdot \log P\left(D = g|\mathcal{C}_V\left(f_S\left(x_i\right)\right)\right), \tag{3}$$

It encourages the identity-specific factors of all images to contain less domain information about their datasets.

As a result, the parameters θ_S of the identity-aware encoder f_S are optimized by jointly minimizing the identity-classification loss and the domain-indistinguishability loss. The overall objective function of identity adversarial learning is:

$$\min_{\theta_S} \mathcal{L}_{id}^{s} + \lambda_1 \mathcal{L}_{indis}^{dom}. \tag{4}$$

Domain Adversarial Learning Block. This block aims to extract the domain-specific factors from person images x_i and exclude identity-specific factors. As it is a dual task to the identity adversarial learning, we can use a similar training schema to learn the parameters of f_V. Here, the identity-classification loss and the domain-indistinguishability loss in the above section are replaced by the domain-classification loss $\mathcal{L}_{dom}^{v}$ and the identity-indistinguishability loss $\mathcal{L}_{indis}^{id}$ respectively. The objective of the domain adversarial learning block is as follows:

$$\min_{\theta_V} \mathcal{L}_{dom}^{v} + \lambda_2 \mathcal{L}_{indis}^{id}. \tag{5}$$

The domain-classification loss $\mathcal{L}_{dom}^{v}$ is defined as,

$$\mathcal{L}_{dom}^{v} = -\sum_{i=1}^{N} \log P(D = d_i|\mathcal{C}_V(f_V(x_i))), \tag{6}$$

where d_i is the domain label of image x_i. The identity-indistinguishability loss $\mathcal{L}_{indis}^{id}$ is similar to Equation 3 but has different number of classes (number of classes for $\mathcal{L}_{indis}^{id}$ is M and G for $\mathcal{L}_{indis}^{dom}$).

Causal Data Augmentation Block. With the disentangled representations learnt from MDDAN, we can bring identity-specific information of x_i to various contexts and backgrounds, which can be seen as identity-preserving data augmentation. For each sample x_i, four different strategies are used to select K domain-specific factors for data augmentation.

- K-Random. For each x_i, we randomly select K different domain-specific factors $\{v_j\}_{j=1}^{K}$ where $j \neq i$.
- K-Hardest. For each x_i, we select K domain-specific factors $\{v_j\}_{j=1}^{K}$ which are the most dissimilar to v_i.
- K-Mixup. For each x_i, we can create more domain-specific factors by mixup [76, 73]. We generate K mixed sample feature $\tilde{v}$ by interpolation of two randomly selected features (a pair $\langle v_j, v_j' \rangle$), denoted by

  $$\tilde{v} = \alpha v_j + (1-\alpha)v_j', \tag{7}$$

  where $\alpha \in [0,1]$ controls the interpolation degree and we empirically set $\alpha = 0.5$ in our experiments.
- K-MixHard. For each x_i, we firstly select K domain-specific factors $\{v_j\}^{K}$ least like v_i and generate K mixed features by randomly interpolating these factors.

Then we perform feature-level augmentation by concatenating the K domain-specfic factors $\{v_j\}_{j=1}^{K}$ with s_i and obtained K new feature factors $\{s_i \oplus v_j\}_{j=1}^{K}$. Recall the assumption of disentanglement that v_j shouldn't related to the pedestrian identity. Hence we use a classifier $\mathcal{C}_C^S$ on these new feature vectors and expect that the classification results should be identity-preserving, which can be explained for learning causal invariance from various interventions on the domain-specific factors. Thus, the identity invariance loss is calculated based on the cross entropy loss as follows.

$$\mathcal{L}_{invar}^{id} = -\sum_{i=1}^{N}\sum_{k=1}^{K} \log P(Y = y_i|\mathcal{C}_C^S(s_i \oplus v_k)), \tag{8}$$

Conversely, we can also perform a domain-preserving data augmentation, where K identity-specific factors are selected to be concatenated with a given domain-specific factor v_i and enforcing the domain classification result is not affected by identity-specific factors. And the domain invariance loss is defined as:

$$\mathcal{L}_{invar}^{dom} = -\sum_{i=1}^{N}\sum_{k=1}^{K} \log P(D = d_i|\mathcal{C}_C^V(v_i \oplus s_k)), \tag{9}$$

where $\mathcal{C}_C^V$ is the domain classifier on the new augmented features.

3.3. Model Summary

Recalling the adversarial learning of identity-aware encoder f_S and domain-aware encoder f_V (Eq. (1) and Eq. (6)), the identity and domain classifiers need to be trained with the inputs of domain-specific and identity-specific factors. The classification losses are calculated in terms of the cross entropy function.

$$\mathcal{L}_{id}^{v} = -\sum_{i=1}^{N} \log P\left(Y = y_i|\mathcal{C}_S\left(f_V\left(x_i\right)\right)\right). \tag{10}$$

$$\mathcal{L}_{dom}^{s} = -\sum_{i=1}^{N} \log P(D = d_i|\mathcal{C}_V(f_S(x_i))). \tag{11}$$

Finally, given the parameters ϕ_S, ϕ_V of classifiers $\mathcal{C}_S, \mathcal{C}_V$, the total loss function is summarized as follows.

$$\min_{\phi_S,\phi_V} \mathcal{L}^s_{id} + \mathcal{L}^s_{dom} + \mathcal{L}^v_{dom} + \mathcal{L}^v_{id}. \tag{12}$$

And the overall loss functions for the other components are defined as

$$\begin{aligned} &\min_{\theta_S,\phi^S_C} \mathcal{L}^s_{id} + \lambda_1 \mathcal{L}^{dom}_{indis} + \lambda_3 \mathcal{L}^{id}_{invar}, \\ &\min_{\theta_V,\phi^V_C} \mathcal{L}^v_{dom} + \lambda_2 \mathcal{L}^{id}_{indis} + \lambda_4 \mathcal{L}^{dom}_{invar}, \end{aligned} \tag{13}$$

where ϕ^S_C, ϕ^V_C are parameters of classifiers $\mathcal{C}^S_C$ and $\mathcal{C}^V_C$ respectively. With the above components, each mini-batch training process is divided into two phases. In the Phase I, the encoders f_S and f_V as well as the augmented data classifiers C^S_C and C^V_C are trained by Eq. (13), while the identity and domain classifiers are fixed. Then in the Phase II, identity and domain classifiers are trained by Eq. (12), while other components are fixed.

3.4. Analysis

Lemma 3.1 *Let $\mathcal{D}_i$ denote one of the source domains, s^τ are identity-specific factors of images from the τ-th domain. Let $p(s^\tau|\mathcal{D}_i)$ be the class-conditional density function of the τ-th domain given domain information $\mathcal{D}_i$. It can be proved that, carrying out the MDDAN will lead to*

$$p(s^\tau|\mathcal{D}_i) = p_\tau(s^\tau), \forall s^\tau \in \mathcal{D}_\tau; i = 1, ..., K. \tag{14}$$

It indicates that in the latent space of identity-specific representation, any image will be invariant to K different domains. Its class-conditional density function value for these domains (e.g., *$p(s^\tau|\mathcal{D}_i)$) is just equal to its prior density function value in its own domain* (e.g., *$p_\tau(s^\tau)$), but not dependents on the domain index i. The proof is shown in the Section A of the supplementary material.*

Lemma 3.2 *From the view of information theory, MDDAN is minimizing the mutual information between identity-specific factors S and domain information D, namely $\min \mathcal{I}(S, D)$. The proof is in the Section B of the supplementary material.*

4. Experiments

4.1. Datasets and Settings

Following [59, 28], we evaluate the DIR-ReID with multiple sources (MS), where source domains cover five large-scale ReID datasets, including CUHK02 [33], CUHK03 [34], Market1501 [82], DukeMTMC-ReID [84], CUHK-SYSU PersonSearch [69]. Details of MS are summarized in Table 1. The unseen test domains are VIPeR [22], PRID [26], QMUL GRID [38] and i-LIDS [68]. We follow the single-shot setting, where the number of probe/gallery images are summarized in Table 2. The evaluation protocols can be found in Section C of the supplementary material. The average rank-k (R-k) accuracy and mean Average Precision (mAP) over 10 random splits are reported based on the evaluation protocol. In this way, we simulate the real-world setting that a ReID model is trained with all the public datasets to evaluate the generalization capability to unseen domains.

Collection	Dataset	IDs	Images
MS	CUHK02	1,816	7,264
	CUHK03	1,467	14,097
	DukeMTMC-Re-Id	1,812	36,411
	Market-1501	1,501	29,419
	CUHK-SYSU	11,934	34,547

Table 1. Training Datasets Statistics. All the images in these datasets, regardless of their original train/test splits, are used for model training.

Dataset	Probe		Gallery	
	Pr. IDs	Pr. Imgs	Ga. IDs	Ga. imgs
PRID	100	100	649	649
GRID	125	125	1025	1,025
VIPeR	316	316	316	316
i-LIDS	60	60	60	60

Table 2. Testing Datasets statistics.

4.2. Implementation Details

Following previous generalizable person ReID methods, we use MobileNetV2 [56] as the domain-specific encoder f_V and use MobileNetV2 with IN layer [37] as identity-specific encoder f_S. Our classifiers $\mathcal{C}_S, \mathcal{C}_V, \mathcal{C}^S_C, \mathcal{C}^V_C$ are simply composed of a single fully-connected layer. Images are resized to 256×128 and the training batch size is set to 128. Random cropping, random flipping, and color jitter are applied as data augmentations. The label smoothing parameter is 0.1. SGD is used to train all the components from scratch with a learning rate 0.02 and momentum 0.9. The training process includes 150 epochs and the learning rate is divided by 10 after 100 epochs. At test time, DIR-ReID only involves identity-specific encoder f_S, which is of a comparable network size to most ReID methods. The tradeoff weights are set to $\lambda_2 = 0.1$ and $\lambda_1 = \lambda_3 = \lambda_4 = 1$ empirically.

4.3. Comparisons Against State-of-the-art

Comparison with Single Domain Methods. Many supervised methods report high performance on large-scale benchmarks, but their performance is still poor on small-scale ones. We select 6 representative models (labeled as ‘S’ in Table 3) trained with data and labels the same as the training splits of target datasets. Although data from the target

Methods	Type	Source	VIPeR (V)				PRID (P)				GRID (G)				i-LIDS (I)			
			R-1	R-5	R-10	mAP	R-1	R-5	R-10	mAP	R-1	R-5	R-10	mAP	R-1	R-5	R-10	mAP
DeepRank [10]	S	Target	38.4	69.2	81.3	-	-	-	-	-	-	-	-	-	-	-	-	-
DNS [78]	S	Target	42.3	71.5	82.9	-	29.8	52.9	66.0	-	-	-	-	-	-	-	-	-
MTDnet [12]	S	Target	47.5	73.1	82.6	-	32.0	51.0	62.0	-	-	-	-	-	58.4	80.4	87.3	-
JLML [78]	S	Target	50.2	**74.2**	84.3	-	-	-	-	-	37.5	61.4	69.4	-	-	-	-	-
SSM [4]	S	Target	53.7	-	**91.5**	-	-	-	-	-	27.2	-	61.2	-	-	-	-	-
SpindleNet [81]	S	Target	**58.3**	74.1	83.2	-	67.0	**89.0**	**89.0**	-	-	-	-	-	66.3	86.6	91.8	-
MMFA [37]	DA	M	38.5	-	-	-	26.8	-	-	-	-	-	-	-	-	-	-	-
MMFA [37]	DA	D	36.3	-	-	-	34.5	-	-	-	-	-	-	-	-	-	-	-
TJ-AIDL [65]	DA	M	38.5	-	-	-	26.8	-	-	-	-	-	-	-	-	-	-	-
TJ-AIDL [65]	DA	D	35.1	-	-	-	34.8	-	-	-	-	-	-	-	-	-	-	-
UDML [49]	DA	Comb	31.5	-	-	-	24.2	-	-	-	-	-	-	-	49.3	-	-	-
SyRI [5]	DA	C3+D	43.0	-	-	-	43.0	-	-	-	-	-	-	-	56.5	-	-	-
AugMining [62]	DG	MS	49.8	70.8	77.0	-	34.3	56.2	65.7	-	46.6	**67.5**	**76.1**	-	76.3	**93.0**	**95.3**	-
DIMN [59]	DG	MS	51.2	70.2	76.0	60.1	39.2	67.0	76.7	52.0	29.3	53.3	65.8	41.1	70.2	89.7	94.5	78.4
DualNorm [28]	DG	MS	53.9	62.5	75.3	58.0	60.4	73.6	84.8	64.9	41.4	47.4	64.7	45.7	74.8	82.0	91.5	78.5
DDAN [9]	DG	MS	56.5	65.6	76.3	60.8	62.9	74.2	85.3	67.5	46.2	55.4	68.0	50.9	**78.0**	85.7	93.2	81.2
DIR-ReID	DG	MS	**58.3**	66.9	77.3	**62.9**	**71.1**	82.4	88.6	**75.6**	**47.8**	51.1	70.5	**52.1**	74.4	83.1	90.2	**78.6**

Table 3. Comparisons against state-of-the-art methods. 'S': single domain, 'DA': domain adaptation (cross-domain), 'DG': domain generalization, 'M': Market1501, 'D': DukeMTMC-ReID, 'comb': the combination of VIPeR, PRID, CUHK01, i-LIDS, and CAVIAR datasets. 'C3': CUHK03, '-': no report. 1^{st} and 2^{ed} highest accuracy are indicated by **blue** and red color.

domain are inaccessible for DIR-ReID, it achieves competitive or better performance on all four benchmarks, which indicates that sufficient source data and our model based on domain invariance learning can alleviate the need for data from the target domain.

Comparison with Domain Adaptation Methods. We also select representative unsupervised domain adaptation methods (labels DA in Table 3) for the comparison. These DA methods are trained with data from the source domain and then adapted to the target domain using unlabeled images from their training splits. Since our model does not perform retraining with unlabeled data, it is more challenging to compare with these DA methods. However, the proposed method still outperforms all of these DA methods by a large margin.

As discussed earlier, supervised methods require labeled target datasets and unsupervised domain adaptation methods require unlabeled images from the target domain, which both need additional training data collections. The expensive manual cost of data collections hinders the ReID models from real-world deployments. DIR-ReID attains superior performance without using any images or labels from the target domain and additional adaptation, which is crucial for practical ReID deployment.

Comparison with DG Methods. Finally, we compare DIR-ReID with existing methods about generalizable person ReID. As far as we know, there are a few publications focusing on person ReID generalization problem, including DIMN [59], DualNorm [28], [62] and DDAN [9]. From the third row in Table 3, the DIR-ReID has achieved the best performance in terms of mAP against other SOTA DG-ReID methods. Although our method falls slightly behind others on the i-LIDS and the GRID datasets in terms of Rank-5 and Rank-10, the DIR-ReID obtains the best Rank-1 performance on three of four datasets. Following [55], we consider the worst-domain accuracy (WDA), which is 32.3% rank 1 accuracy for AugMining, 29.3% rank 1 accuracy for DIMN, 41.4% rank 1 accuracy for DualNorm and 46.2% rank 1 accuracy for DDAN. Compared to them, the DIR-ReID has a superior WDA value, which is 47.8% rank 1 accuracy when tested on the GRID dataset.

	R-1	R-3	R-5	R-10
Baseline (DualNorm [28])	34.4	40.9	46.3	54.7
w/ Dual DANN [19]	35.4	42.0	46.8	54.4
Improvements↑	1.0	1.1	0.5	-0.3
w/ MDDAN	36.6	45.8	51.6	56.2
Improvements↑	2.2	4.9	5.3	1.5
w/ CDA	40.2	53.2	57.2	63.3
Improvements↑	5.8	12.3	10.9	8.6

Table 4. Ablation studies on three different blocks: dual GRL block, our MDDAN block and the CDA block. The reported metric are the rank-1, rank-5, rank-10 accuracies of the GRID dataset. Models here are all trained on three datasets *i.e.* Market-1501, CUHK02 and CUHK03. The improvements are the difference from the baseline.

Overall, the DIR-ReID has achieved competitive performance in the challenging setting of generalizable person ReID, which verify that the identity-specific factors can be disentangled effectively by the proposed dual adversarial learning and casual data augmentation.

4.4. Ablation Studies

There are two main components in the proposed DIR-ReID: the MDDAN block and the CDA block. Here, we firstly analyze the effectiveness of the blocks respectively, then demonstrate their contributions to the final performance of the whole DIR-ReID model.

Analysis of MDDAN. The superiority of MDDAN is verified by comparison with the dual DANN [19] block. The latter means inserting the GRL layers between f_S, C_V

and f_V, C_S. Simultaneously, the dual GRL block replaces the maximum entropy loss in MDDAN with the maximum misclassification loss. Results are shown in Table 8. The dual DANN method only has a slight improvement in comparison with the baseline, while our MDDAN block leads to significant improvements.

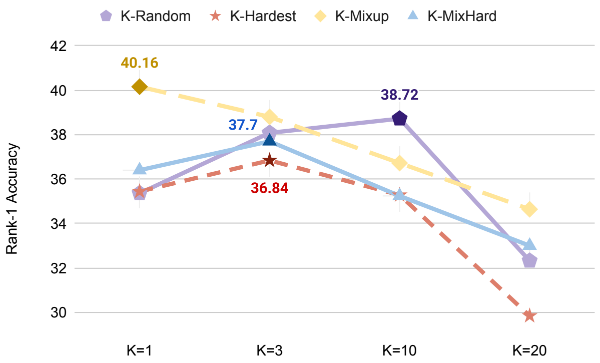


Figure 4. Ablation study on CDA block. The metric is the rank-1 accuracy on the GRID dataset. Considering the expensive cost of training with five datasets, all the models here are trained on three datasets, *i.e.* Market-1501, CUHK02 and CUHK03.

Analysis of Methods for Causal Data augmentation. We compare four different implementations of CDA and the results are shown in Figure 6. K-Random is the simplest method while it works well, attaining 38.72% rank 1 accuracy when $K = 10$. Surprisingly, K-Hardest is no better than K-Random. It may because that these selected hardest vectors may hinder the convergence speed of our optimization process. K-Mixup outperforms other methods and attains 40.17% rank 1 accuracy, which verified the importance of mixup for data augmentation. However, as we increase the value of K, the performance will not be improved. When we combine K-Hardest and mixup, K-MixHard has a slightly higher performance than K-Hardest, while it still has a large margin compared to $K-$MixUp.

Ablation Study of Different Blocks. To evaluate the contribution of each component, we gradually add the MDDAN and the CDA to the baseline, and the overall ablation studies are reported in Table 8. The MDDAN improves the rank-1 accuracy from 34.4% to 36.6%. The results in rank-3, rank-5, rank-10 are consistently improved, which validate that the MDDAN removes some of the domain-specific information from our identity-specific representations and yields consistent generalization performance improvements. The CDA provides greater improvement gains (5.8% higher rank-1 accuracy and more than 8% point higher in other metrics). It validates that CDA can exclude domain-specific information efficiently. Results of ablation studies on other test datasets are reported in Table 5, which also verifies the efficiency of the DIR-ReID. Unfortunately, the DIR-ReID seems not to work well on i-LIDS dataset, we will explore the reason and make improvements in future work.

	PRID	VIPeR	i-LIDS
Baseline (DualNorm [28])	38.9	54.0	64.3
w/ Dual DANN [19]	41.0	53.3	66.2
Improvements↑	2.1	-0.7	1.9
w/ MDDAN	42.7	56.5	65.7
Improvements↑	3.8	2.5	1.4
w/ CDA	43.8	61.4	64.2
Improvements↑	4.9	7.4	-0.1

Table 5. Ablation studies on three different blocks: dual GRL block, our MDDAN block and the CDA block. The metric is the rank-1 accuracy. Models here are all trained on three datasets, including Market-1501, CUHK02 and CUHK03. The improvements denote the difference from the baseline.

We also conduct additional ablation studies and disentanglement experiments on rotated MNIST. The dataset and results in Section D of the supplementary material.

5. Conclusions and Future Work

We propose a novel generalizable person ReID framework based on disentanglement and augmentation from a causal invariant learning perspective. Specifically, a MDDAN block is proposed to disentangle identity-specific and domain-specific factors from multi-source ReID training data. We then propose CDA block to learn invariant identity-specific features. The comprehensive experimental results show that DIR-ReID achieves state-of-the-art performance. We believe DIR-ReID may shed light on the development of new paradigms for DG beyond the person ReID.

To upgrade DIR-ReID, we can improve the model performance with other tricks *i.e.* triplet loss [40], Maximum Mean Discrepancy (MMD) [23] regularization. One promising way is to generate realistic images from concatenated vectors. The augmented feature vectors are guided by a reconstruction loss, which will further improve the disentanglement of identity-specific and domain-specific factors. These generated images can also be used for interpretability and augmenting the training set. Besides, we will seek other methods for better disentanglement performance such as replacing the multi-domain adversarial learning with mutual information minimization [6] or f-divergence [46] maximization. We believe that the methodology developed in this paper can shed light on the study of new paradigms for domain generalization beyond the person ReID.

References

[1] Kartik Ahuja, Karthikeyan Shanmugam, Kush Varshney, and Amit Dhurandhar. Invariant risk minimization games. In

International Conference on Machine Learning, pages 145–155. PMLR, 2020. 3

[2] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization, 2020. 3

[3] Yuval Atzmon, Felix Kreuk, Uri Shalit, and Gal Chechik. A causal view of compositional zero-shot recognition. *Advances in Neural Information Processing Systems (NeurIPS)*, 33, 2020. 3

[4] Song Bai, Xiang Bai, and Qi Tian. Scalable person re-identification on supervised smoothed manifold. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 7

[5] Slawomir Bak, Peter Carr, and Jean-Francois Lalonde. Domain adaptation through synthesis for unsupervised person re-identification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018. 7

[6] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and R Devon Hjelm. Mine: mutual information neural estimation. *arXiv preprint arXiv:1801.04062*, 2018. 8

[7] Michel Besserve, Arash Mehrjou, Rémy Sun, and Bernhard Schölkopf. Counterfactuals uncover the modular structure of deep generative models. In *International Conference on Learning Representations (ICLR)*, 2019. 3

[8] Ruichu Cai, Zijian Li, Pengfei Wei, Jie Qiao, Kun Zhang, and Zhifeng Hao. Learning disentangled semantic representation for domain adaptation. In *IJCAI: proceedings of the conference*, volume 2019, page 2060. NIH Public Access, 2019. 3, 4

[9] Peixian Chen, Pingyang Dai, Jianzhuang Liu, Feng Zheng, Qi Tian, and Rongrong Ji. Dual distribution alignment network for generalizable person re-identification. *Thirty-Fifth AAAI Conference on Artificial Intelligence*, 2021. 7

[10] Shi-Zhe Chen, Chun-Chao Guo, and Jian-Huang Lai. Deep ranking for person re-identification via joint representation learning. *IEEE Transactions on Image Processing*, 25(5):2353–2367, 2016. 7

[11] Weihua Chen, Xiaotang Chen, Jianguo Zhang, and Kaiqi Huang. Beyond triplet loss: a deep quadruplet network for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 403–412, 2017. 2

[12] Weihua Chen, Xiaotang Chen, Jianguo Zhang, and Kaiqi Huang. A multi-task deep network for person re-identification. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, pages 3988–3994, 2017. 7

[13] Weijian Deng, Liang Zheng, Qixiang Ye, Guoliang Kang, Yi Yang, and Jianbin Jiao. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 994–1003, 2018. 2

[14] Shengyong Ding, Liang Lin, Guangrun Wang, and Hongyang Chao. Deep feature learning with relative distance comparison for person re-identification. *Pattern Recognition*, 48(10):2993–3003, 2015. 2

[15] Zhang Dong, Zhang Hanwang, Tang Jinhui, Hua Xiansheng, and Sun Qianru. Causal intervention for weakly supervised semantic segmentation. In *NeurIPS*, 2020. 3

[16] Chanho Eom and Bumsub Ham. Learning disentangled representation for robust person re-identification. In *Advances in Neural Information Processing Systems*, pages 5298–5309, 2019. 1

[17] Hehe Fan, Liang Zheng, Chenggang Yan, and Yi Yang. Unsupervised person re-identification: Clustering and fine-tuning. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 14(4):1–18, 2018. 2

[18] Yang Fu, Yunchao Wei, Guanshuo Wang, Yuqian Zhou, Honghui Shi, and Thomas S Huang. Self-similarity grouping: A simple unsupervised cross domain adaptation approach for person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6112–6121, 2019. 2

[19] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015. 4, 7, 8, 14

[20] Yixiao Ge, Dapeng Chen, and Hongsheng Li. Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification. *arXiv preprint arXiv:2001.01526*, 2020. 2

[21] Muhammad Ghifary, W. Bastiaan Kleijn, Mengjie Zhang, and David Balduzzi. Domain generalization for object recognition with multi-task autoencoders. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015. 13

[22] D. Gray, S. Brennan, and H. Tao. Evaluating Appearance Models for Recognition, Reacquisition, and Tracking. *Proc. IEEE International Workshop on Performance Evaluation for Tracking and Surveillance (PETS)*, 2007. 6, 13

[23] Arthur Gretton, Kenji Fukumizu, Zaid Harchaoui, and Bharath K Sriperumbudur. A fast, consistent kernel two-sample test. In *NIPS*, volume 23, pages 673–681, 2009. 8

[24] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*, 2020. 2, 3

[25] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. *arXiv preprint arXiv:2006.16241*, 2020. 2

[26] Martin Hirzer, Csaba Beleznai, Peter M Roth, and Horst Bischof. Person re-identification by descriptive and discriminative classification. In *Scandinavian conference on Image analysis*, pages 91–102. Springer, 2011. 6, 13

[27] Yukun Huang, Zheng-Jun Zha, Xueyang Fu, Richang Hong, and Liang Li. Real-world person re-identification via degradation invariance learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14084–14094, 2020. 2

[28] Jieru Jia, Qiuqi Ruan, and Timothy M Hospedales. Frustratingly easy person re-identification: Generalizing person

re-id in practice. *arXiv preprint arXiv:1905.03422*, 2019. 2, 3, 6, 7, 8

[29] Xin Jin, Cuiling Lan, Wenjun Zeng, Zhibo Chen, and Li Zhang. Style normalization and restitution for generalizable person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3143–3152, 2020. 2, 3

[30] Srikrishna Karanam, Yang Li, and Richard J Radke. Person re-identification with discriminatively trained viewpoint invariant dictionaries. In *Proceedings of the IEEE international conference on computer vision*, pages 4516–4524, 2015. 2

[31] Martin Koestinger, Martin Hirzer, Paul Wohlhart, Peter M Roth, and Horst Bischof. Large scale metric learning from equivalence constraints. In *2012 IEEE conference on computer vision and pattern recognition*, pages 2288–2295. IEEE, 2012. 2

[32] Dangwei Li, Xiaotang Chen, Zhang Zhang, and Kaiqi Huang. Learning deep context-aware features over body and latent parts for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 384–393, 2017. 2

[33] Wei Li and Xiaogang Wang. Locally aligned feature transforms across views. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2013. 6

[34] Wei Li, Rui Zhao, Tong Xiao, and Xiaogang Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014. 6

[35] Shengcai Liao, Yang Hu, Xiangyu Zhu, and Stan Z Li. Person re-identification by local maximal occurrence representation and metric learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2197–2206, 2015. 2

[36] Shan Lin, Chang-Tsun Li, and Alex C Kot. Multi-domain adversarial feature generalization for person re-identification. *arXiv preprint arXiv:2011.12563*, 2020. 2, 3

[37] Shan Lin, Haoliang Li, Chang Tsun Li, and Alex Chichung Kot. Multi-task mid-level feature alignment network for unsupervised cross-dataset person re-identification. In *29th British Machine Vision Conference, BMVC 2018*, 2018. 6, 7

[38] Chunxiao Liu, Shaogang Gong, Chen Change Loy, and Xinggang Lin. Person re-identification: What features are important? In *European Conference on Computer Vision*, pages 391–401. Springer, 2012. 6, 13

[39] Chang Liu, Xinwei Sun, Jindong Wang, Tao Li, Tao Qin, Wei Chen, and Tie-Yan Liu. Learning causal semantic representation for out-of-distribution prediction. *arXiv preprint arXiv:2011.01681*, 2020. 3

[40] Hao Liu, Jiashi Feng, Meibin Qi, Jianguo Jiang, and Shuicheng Yan. End-to-end comparative attention networks for person re-identification. *IEEE Transactions on Image Processing (TIP)*, 26(7):3492–3506, 2017. 8

[41] Chuanchen Luo, Chunfeng Song, and Zhaoxiang Zhang. Generalizing person re-identification by camera-aware invariance learning and cross-domain mixup. In *European Conference on Computer Vision (ECCV)*, 2020. 1, 2

[42] Sara Magliacane, Thijs van Ommen, Tom Claassen, Stephan Bongers, Philip Versteeg, and Joris M Mooij. Domain adaptation by using causal inference to predict invariant conditional distributions. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 10846–10856, 2018. 3

[43] Tetsu Matsukawa, Takahiro Okabe, Einoshin Suzuki, and Yoichi Sato. Hierarchical gaussian descriptor for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1363–1372, 2016. 2

[44] Jovana Mitrovic, Brian McWilliams, Jacob Walker, Lars Buesing, and Charles Blundell. Representation learning via invariant causal mechanisms. *International Conference on Learning Representations (ICLR)*, 2020. 3

[45] Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. When does label smoothing help? In *Advances in Neural Information Processing Systems*, pages 4694–4703, 2019. 4

[46] Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-gan: Training generative neural samplers using variational divergence minimization. *arXiv preprint arXiv:1606.00709*, 2016. 8

[47] Giambattista Parascandolo, Niki Kilbertus, Mateo Rojas-Carulla, and Bernhard Schölkopf. Learning independent causal mechanisms. In *ICML*, 2018. 3

[48] Judea Pearl et al. Models, reasoning and inference. *Cambridge, UK: CambridgeUniversityPress*, 2000. 3

[49] Peixi Peng, Tao Xiang, Yaowei Wang, Massimiliano Pontil, Shaogang Gong, Tiejun Huang, and Yonghong Tian. Unsupervised cross-dataset transfer learning for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 2, 7

[50] Xingchao Peng, Yichen Li, and Saenko Kate. Domain2vec: Domain embedding for unsupervised domain adaptation. *European conference on computer vision (ECCV 2020)*, 2020. 2

[51] Jiaxin Qi, Yulei Niu, Jianqiang Huang, and Hanwang Zhang. Two causal principles for improving visual dialog. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 3

[52] Lei Qi, Lei Wang, Jing Huo, Luping Zhou, Yinghuan Shi, and Yang Gao. A novel unsupervised camera-aware domain adaptation framework for person re-identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 1, 12

[53] Mateo Rojas-Carulla, Bernhard Schölkopf, Richard Turner, and Jonas Peters. Invariant models for causal transfer learning. *The Journal of Machine Learning Research (JMLR)*, 19(1):1309–1342, 2018. 3

[54] Elan Rosenfeld, Pradeep Ravikumar, and Andrej Risteski. The risks of invariant risk minimization, 2020. 3

[55] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for

group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019. 7

[56] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018. 6

[57] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Towards causal representation learning. *arXiv preprint arXiv:2102.11107*, 2021. 3

[58] Chunfeng Song, Yan Huang, Wanli Ouyang, and Liang Wang. Mask-guided contrastive attention model for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1179–1188, 2018. 2

[59] Jifei Song, Yongxin Yang, Yi-Zhe Song, Tao Xiang, and Timothy M. Hospedales. Generalizable person re-identification by domain-invariant mapping network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 1, 2, 3, 6, 7

[60] Xinwei Sun, Botong Wu, Chang Liu, Xiangyu Zheng, Wei Chen, Tao Qin, and Tie-yan Liu. Latent causal invariant model. *arXiv preprint arXiv:2011.02203*, 2020. 3

[61] Yifan Sun, Liang Zheng, Yi Yang, Qi Tian, and Shengjin Wang. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018. 2

[62] Masato Tamura and Tomokazu Murakami. Augmented hard example mining for generalizable person re-identification. *arXiv preprint arXiv:1910.05280*, 2019. 2, 7

[63] Kaihua Tang, Jianqiang Huang, and Hanwang Zhang. Long-tailed classification by keeping the good and removing the bad momentum causal effect. In *NeurIPS*, 2020. 3

[64] Jingya Wang, Xiatian Zhu, Shaogang Gong, and Wei Li. Transferable joint attribute-identity deep learning for unsupervised person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2275–2284, 2018. 2

[65] Jingya Wang, Xiatian Zhu, Shaogang Gong, and Wei Li. Transferable joint attribute-identity deep learning for unsupervised person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 7

[66] Tan Wang, Jianqiang Huang, Hanwang Zhang, and Qianru Sun. Visual commonsense r-cnn. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10760–10770, 2020. 3

[67] Yanan Wang, Shengcai Liao, and Ling Shao. Surpassing real-world source training data: Random 3d characters for generalizable person re-identification. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 3422–3430, 2020. 2, 3

[68] Zheng Wei-Shi, Gong Shaogang, and Xiang Tao. Associating groups of people. In *Proceedings of the British Machine Vision Conference*, pages 23–1, 2009. 6, 13

[69] Tong Xiao, Shuang Li, Bochao Wang, Liang Lin, and Xiaogang Wang. End-to-end deep learning for person search. *arXiv preprint arXiv:1604.01850*, 2(2), 2016. 6

[70] Xun Yang, Meng Wang, and Dacheng Tao. Person re-identification with metric learning using privileged information. *IEEE Transactions on Image Processing*, 27(2):791–805, 2017. 2

[71] Xu Yang, Hanwang Zhang, and Jianfei Cai. Deconfounded image captioning: A causal retrospect. *arXiv preprint arXiv:2003.03923*, 2020. 3

[72] Yang Yang, Jimei Yang, Junjie Yan, Shengcai Liao, Dong Yi, and Stan Z Li. Salient color names for person re-identification. In *European conference on computer vision*, pages 536–551. Springer, 2014. 2

[73] Mang Ye, Jianbing Shen, Xu Zhang, Pong C Yuen, and Shih-Fu Chang. Augmentation invariant and instance spreading feature for softmax embedding. *IEEE transactions on pattern analysis and machine intelligence*, 2020. 5

[74] Zhongqi Yue, Hanwang Zhang, Qianru Sun, and Xian-Sheng Hua. Interventional few-shot learning. In *NeurIPS*, 2020. 3

[75] Yunpeng Zhai, Qixiang Ye, Shijian Lu, Mengxi Jia, Rongrong Ji, and Yonghong Tian. Multiple expert brainstorming for domain adaptive person re-identification. *arXiv preprint arXiv:2007.01546*, 2020. 2

[76] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations (ICLR)*, 2018. 5

[77] Ling Zhang, Xiaosong Wang, Dong Yang, Thomas Sanford, Stephanie Harmon, Baris Turkbey, Holger Roth, Andriy Myronenko, Daguang Xu, and Ziyue Xu. When unseen domain generalization is unnecessary? rethinking data augmentation. *arXiv preprint arXiv:1906.03347*, 2019. 2, 3

[78] Li Zhang, Tao Xiang, and Shaogang Gong. Learning a discriminative null space for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 7

[79] Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, and Zhibo Chen. Densely semantically aligned person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 667–676, 2019. 2

[80] Zhizheng Zhang, Cuiling Lan, Wenjun Zeng, Xin Jin, and Zhibo Chen. Relation-aware global attention for person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1

[81] Haiyu Zhao, Maoqing Tian, Shuyang Sun, Jing Shao, Junjie Yan, Shuai Yi, Xiaogang Wang, and Xiaoou Tang. Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 7

[82] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015. 6

[83] Zhedong Zheng, Xiaodong Yang, Zhiding Yu, Liang Zheng, Yi Yang, and Jan Kautz. Joint discriminative and generative learning for person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 1, 3

[84] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 6

A. Theoretical analysis of the Multi-Domain Disentangled Adversarial Neural Network (MDDAN)

Here we adapt from [52] and prove the Lemma 1 in the original manuscript. All the analyses are conducted in the shared ID-specific representation space. We first prove the following lemma.

Lemma A.1 *When MDDAN is carried out and the uncertainty is maximized, given any ID-specific representation* $\mathbf{s}^{\tau}$ *of an image from domain* $\mathcal{D}_{\tau}$*, the conditional probability of any* $\mathcal{D}_i$ *given* $\mathbf{s}^{\tau}$ *will equal* $1/K$*, namely for any* $i = 1, ..., K$*, we have* $p(\mathcal{D}_i|\mathbf{s}^{\tau}) = 1/K$*.*

Proof A.1 *Here we slightly abuse the notation: omitting the domain classifier* $\mathcal{C}_V$ *and* $p(k|\mathbf{s}^{\tau})$ *is equivalent to* $p(\mathcal{D}_k|\mathbf{s}^{\tau})$*. MDDAN is defined as follows.*

$$\begin{aligned} &\max -\sum_{k=1}^{K} p(k|\mathbf{s}^{\tau}) \log p(k|\mathbf{s}^{\tau}), \\ &s.t. \sum_{k=1}^{K} p(k|\mathbf{s}^{\tau}) = 1; p(k|\mathbf{s}^{\tau}) \geq 0, \forall k. \end{aligned} \tag{15}$$

Let $\mathcal{H}(p_k) = p_k \log p_k$*, where* $p_k = p(k|\mathbf{s}^{\tau})$*. We simplify the above optimization problem as*

$$\min \sum_{k=1}^{K} \mathcal{H}(p_k), \; s.t. \sum_{k=1}^{K} p_k = 1; p_k \geq 0, \forall k, \tag{16}$$

where $\mathcal{H}(p_k)$ *is convex as we have* $\mathcal{I}''(p_k) = \frac{1}{p_k \ln 2} > 0$*. The sum of convex functions* $\sum_{k=1}^{K} \mathcal{H}(p_k)$ *is also a convex function. Namely, this problem is a convex optimization problem. To prove the lemma, it is now equivalent to show that the minimum value of this convex optimization problem is obtained when* $p_1 = p_2 = ... = p_K = 1/K$*.*

We use the Lagrange multiplier method to solve the problem. The Lagrange function is

$$L(p, \lambda) = \sum_{k=1}^{K} p_k \log p_k + \lambda(\sum_{k=1}^{K} p_k - 1)). \tag{17}$$

We take partial derivatives for each p_k *and get*

$$\frac{\partial L(p, \lambda)}{\partial p_k} = \log p_k + \frac{1}{\ln 2} + \lambda = 0. \tag{18}$$

Then we have $p_k = 2^{-\lambda - 1/\ln 2}$*. As* $\sum_{k=1}^{K} p_k = 1$*, we have* $K * 2^{-\lambda - 1/\ln 2} = 1$*, and thus* $p_k = 1/K, i = 1, 2, ..., K$*. Since the local minimum of the convex function is the global minimum, when* $p_1 = p_2 = ... = p_K = 1/K$*,* $\sum_{k=1}^{K} \mathcal{H}(p_k)$ *achieves the minimum value,* $\log \frac{1}{K}$*. In other words, when MDDAN is carried out and achieves the maximum uncertainty, for all* $i = 1, ..., K$*, we have* $p(\mathcal{D}_i|\mathbf{s}^{\tau}) = 1/K$*.*

Now we are ready to prove Lemma 1. we can calculate any domain $\mathcal{D}_i$'s conditional probability given $\mathbf{s}^\tau$, which is

$$p(\mathcal{D}_i|\mathbf{s}^\tau) = \frac{p(\mathbf{s}^\tau|\mathcal{D}_i)p(\mathcal{D}_i)}{p_\tau(\mathbf{s}^\tau)}, \forall \mathbf{s}^\tau \in \mathcal{D}_\tau; i = 1, ..., K, \tag{19}$$

where $p(\mathbf{s}^\tau|\mathcal{D}_i)$ denotes the conditional probability of $\mathbf{s}^\tau$ given domain information $\mathcal{D}_i$, $p_\tau(\mathbf{s}^\tau)$ denotes the probability function of the ID-specific representation in its domain $\mathcal{D}_\tau$, and $p(\mathcal{D}_i)$ is the priori probability of domain classes. Without generality, we set equal priori probability for each domain, namely $p(\mathcal{D}_i) = 1/K$. Further, from the Lemma A we know that, optimizing MDDAN leads to $p(\mathcal{D}_i|\mathbf{s}^\tau) = 1/K$ for all $i = 1, ..., K$. Hence, Eq.(19) becomes

$$\begin{aligned} 1/K = &\frac{p(\mathbf{s}^\tau|\mathcal{D}_i)1/K}{p_\tau(\mathbf{s}^\tau)}, \forall \mathbf{s}^\tau \in \mathcal{D}_\tau; i = 1, ..., K, \\ \Rightarrow & p(\mathbf{s}^\tau|\mathcal{D}_i) = p_\tau(\mathbf{s}^\tau), \forall \mathbf{s}^\tau \in \mathcal{D}_\tau; i = 1, ..., K, \end{aligned} \tag{20}$$

thus completes the proof.

B. Understanding MDDAN by information theory

Minimizing the mutual information between the ID-specific factors S and the domain information D is defined as

$$\begin{aligned} \min \mathcal{I}(S, D) &= \min \mathcal{H}(D) - \mathcal{H}(D|S) \\ &= \min -\mathcal{H}(D|S) \\ &= \max \mathcal{H}(D|S) \end{aligned} \tag{21}$$

The second line is derived since the entropy of domain distribution $\mathcal{H}(D)$ is not related to our optimization. Namely, our MDDAN is essentially minimizing the mutual information between the ID-specific factors and the domain information.

C. Evaluation Protocols.

Evaluation protocols are as follows.

GRID [38] contains 250 probe images and 250 true match images of the probes in the gallery. Besides, there are a total of 775 additional images that do not belong to any of the probes. We randomly take out 125 probe images. The remaining 125 probe images and $775 + 250$ images in the gallery are used for testing.

i-LIDS [68] has two versions, images and sequences. The former is used in our experiments. It involves 300 different pedestrian pairs observed across two disjoint camera views 1 and 2 in public open space. We randomly select 60 pedestrian pairs, two images per pair are randomly selected as probe image and gallery image respectively.

block	details
1	Conv2d(32, 5), BatchNorm2d, ReLU
2	MaxPool2d(2, 2)
3	Conv2d(64, 5), BatchNorm2d, ReLU
4	MaxPool2d(2, 2)
5	Linear(2)

Table 6. Architecture for encoders f_S, f_V. The parameters for Conv2d are output channels and kernel size. Theparameters for MaxPool2d are kernel size and stride. The parameter for Linear is output features.

PRID2011 [26] has single-shot and multi-shot versions. We use the former in our experiments. The single-shot version has two camera views A and B, which capture 385 and 749 pedestrians respectively. Only 200 pedestrians appear in both views. During the evaluation, 100 randomly identities presented in both views are selected, the remaining 100 identities in view A constitute probe set and the remaining 649 identities in view B constitute gallery set.

VIPeR [22] contains 632 pedestrian image pairs. Each pair contains two images of the same individual seen from different cameras 1 and 2. Each image pair was taken from an arbitrary viewpoint under varying illumination conditions. To compare to other methods, we randomly select half of these identities from camera 1 as probe images and their matched images in 2 as gallery images.

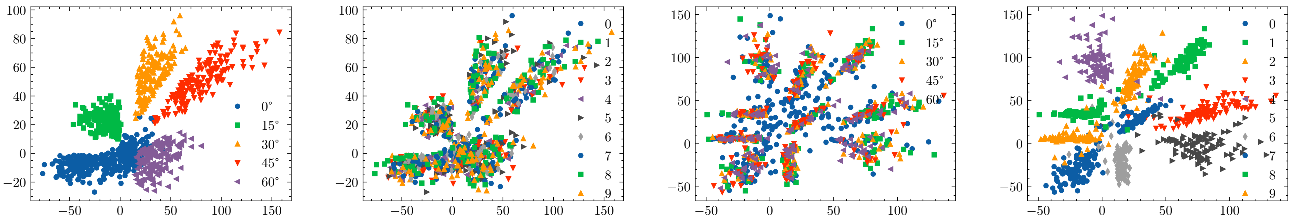
D. Ablation studies on Rotated MNIST.

D.1. Dataset and Setting

To verify whether the capability of DIR-ReID model to disentangling s and v, we first construct rotated MNIST datasets following [21]. 100 images per class (10 classes totally) are randomly sampled from the MNIST training dataset, which is denoted by $\mathcal{M}_{0^\circ}$. We then rotated the images in $\mathcal{M}_{0^\circ}$ by $15, 30, 45$, and 60 degrees, creating four additional domains. To plot the latent subspaces directly without applying dimensionality reduction, we restrict the size of each latent space to 2 dimensions. In experiments, we train a simplified version of the MDDAN with an additional reconstruction loss, which is enough to attain a encouraging result.

D.2. Architecture and Implementation Details

The architectures of the encoders, classifiers are displayed in Table 6, Table 7 respectively. All methods re trained for 500 epochs and the batch size is set to be 100. Adam is used to train all the components from scratch with learning rate as 0.001. We also use warmup to linearly increase the learning rate from 0 to 0.001 during the first 100 epochs of training.



(a) (b) (c) (d)

Figure 5. 2D embeddings of all two latent subspaces. (a,b): v encoded by f_V. (c,d): s encoded by f_S. (a,c) are colored according to their domains and (b,d) are colored according to their classes.

block	details
For $\mathcal{C}_V$	ReLU,Linear(5)
For $\mathcal{C}_S$	ReLU,Linear(10)
For $\mathcal{C}_C^S$	ReLU,Linear(5)
For $\mathcal{C}_C^v$	ReLU,Linear(5)

Table 7. Architecture of classifiers for ID-specific factors, Domain-Specific factors, and concatenated vectors. The parameter for Linear is output features.

	Test Accuracy on $\mathcal{M}_{75^\circ}$
Baseline	46.4
w/ Dual DANN [19]	53.5
Improvements↑	7.1
w/ MDDAN	58.2
Improvements↑	11.8
w/ CDA	61.9
Improvements↑	15.5

Table 8. Ablation study of three different blocks: dual GRL block, our MDDAN block and the CDA block. The reported validation metric is the accuracy of the $\mathcal{M}_{75^\circ}$ dataset. The baseline is only using the encoder f_S and classifier $\mathcal{C}_S$.

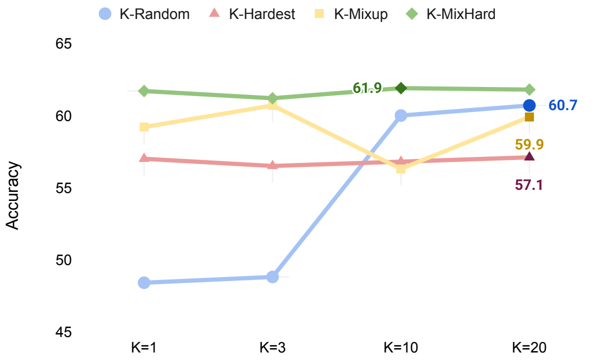


Figure 6. Ablation study of causal data augmentation methods. The reported validation metric is the test accuracy of the $\mathcal{M}_{70^\circ}$ dataset.

D.3. Additional Experimental Results

Analysis of MDDAN. As shown in Table 8, directly applying the dual GRL block benefits the generalization ability, while the proposed MDDAN improve the test accuracy on $\mathcal{M}_{75^\circ}$ dataset by a even more large margin, which is 11.8% points.

Analysis of Methods for Causal Data Augmentation. The comparison results are shown in Figure 6. Different from CDA for Re-ID, here K-MixHard attains the most superior performance, which is 61.9% test accuracy. In general, $K-$Random can attain an encouraging performance gain and we can choose other kinds of methods and parameters based on the difficulty of the dataset.

Ablation Study of Different Blocks. By adding the multi-domain disentangled block and the causal data augmentation block successively, we improve the generalization accuracy from 46.4% to 58.2% and 61.9% respectively (Table 8), showing the effectiveness of the proposed model again.

Disentanglement on Rotated MNIST. To bring the ability of causal reasoning into DDARe-ID, the most important thing is to distinguish the ID-specific factors. The disentanglement results are demonstrated in Figure 5. We can find a correlation between the rotation angle (domain labels) and v in Figure 5(a), five domains are clustered into five distinct clusters, while in Figure 5(b) no clustering is visible, namely there is very weak correlation between v and class labels. By contrast, in Figure 5(c) no clustering is visible according to the rotation angle. But Figure 5(d) shows ten distinct clusters, where each cluster corresponds to a class. From these initial qualitative results, we conclude that the dual MDDAN is disentangling the information contained in x as intended. v only contains domain information and s only contains identity information. The great disentangle results ensure the causal data augmentation can be carried out effectively.